Reading TM5-MP meteorological data using the XIOS input/output server

Jacob van Peet

33rd International TM5 Meeting

IUP/LAMOS/Universität Bremen/online, 19&20 December 2022

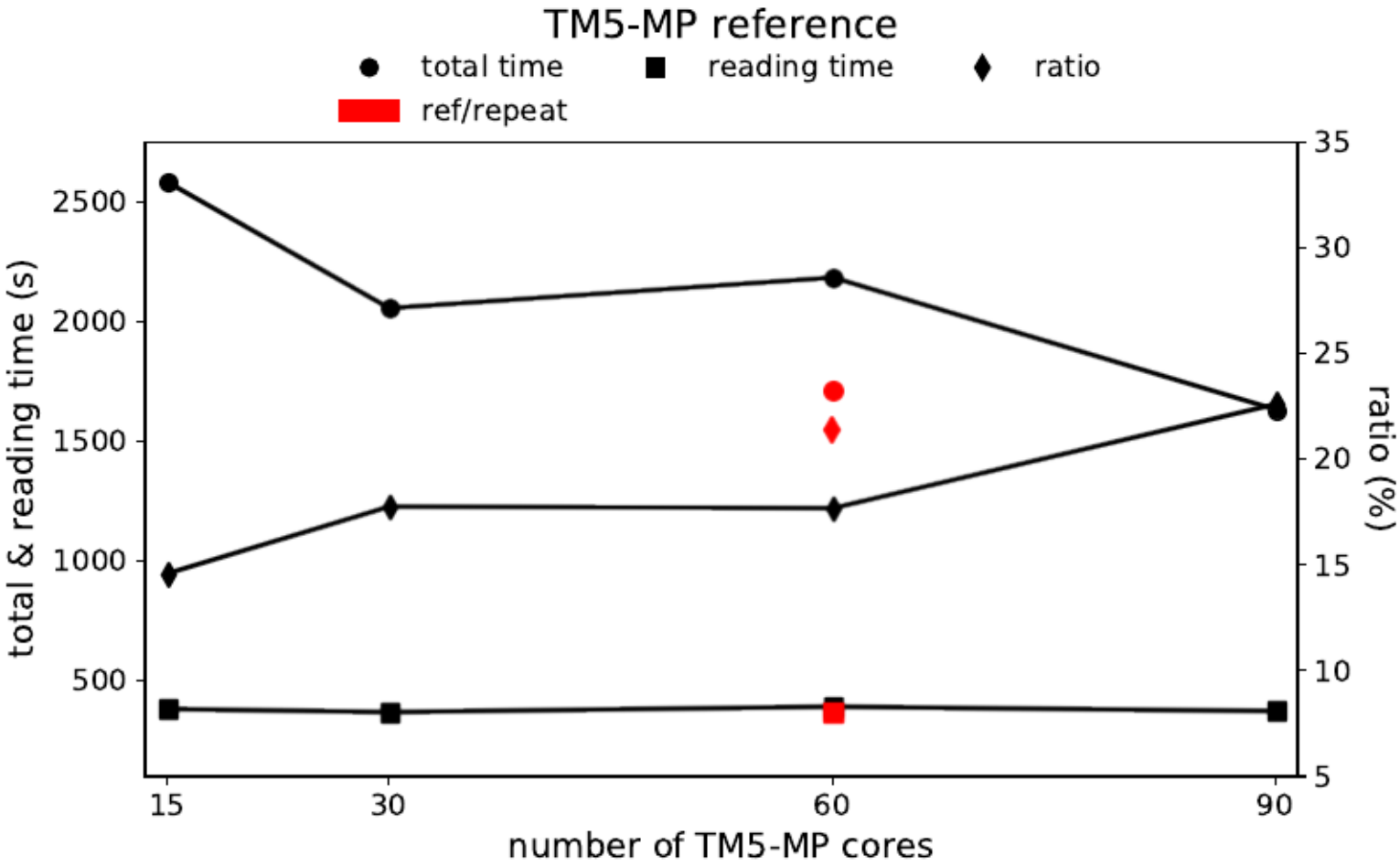
Overview

- Introduction
- Problem visualisation
- XIOS explained
- TM5-MP / XIOS interface
- TM5-MP / XIOS test results
- Conclusion / Outlook

Introduction

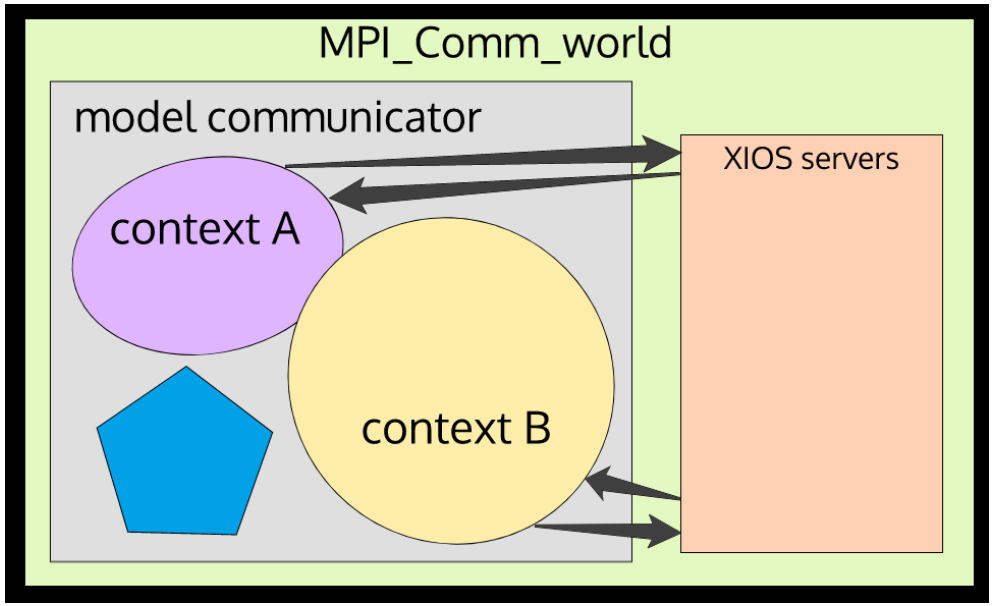
- The more cores you use to run TM5-MP, reading the meteorological data required to run the model will take relatively more time
- Solution: use the XML Input Output Server (XIOS) for reading meteo data
  - stand-alone program that runs next to your model which is dedicated to reading (and) writing data
  - while reading data, the model can continue with its calculations
- We expect that reading the data with XIOS takes less time
  - makes it easier to scale the model to more cores
  - which enables running the model on a higher resolution
- Ultimately we want to run TM5-MP globally on 1x1 degree for methane inversions in the CAMS project

Problem visualisation



XIOS explained

- Configuration is XML based, but there's also a Fortran interface
- Built around the concept of a "context"
- Each context has an associated calendar and can contain one or more files with one or more fields
  - The calendar can only progress forward in time
  - All data for a meteo field for the whole model run must be in a single file



from XIOS TUTORIAL : CEA/LSCE – IPSL

Hello world in XIOS

```
<xios>
  <context id="hello_word" >

    <axis_definition>
      <axis id="vertical_axis" n_glo="100" />
    </axis_definition>

    <domain_definition>
      <domain id="horizontal_domain" ni_glo="100" nj_glo="100" />
    </domain_definition>

    <grid_definition>
      <grid id="grid_3d">
       < domain domain_ref="horizontal_domain" >
       < axis   axis_ref="vertical_axis" >
    </grid_definition>

    <field_definition >
      <field id="a_field" operation="average" grid_ref="grid_3d" />
    </field_definition>

    <file_definition type="one_file" output_freq="1d" enabled=".TRUE.">
      <file id="output" name="hello_world" output_freq="1d">
        <field field_ref="a_field" />
      </file>
    </file_definition>

  </context>
</xios>
```

```
SUBROUTINE hello_world(rank,size)
 USE xios
 IMPLICIT NONE
     INTEGER :: rank, size, timestep
 TYPE(xios_duration)     :: dtime
 DOUBLE PRECISION,ALLOCATABLE :: lon(:,:), lat(:,:), field (:,:)
 INTEGER :: ni, nj, ibegin, jbegin

 CALL xios_initialize("client", return_comm=comm)
 CALL xios_context_initialize("hello_world", comm)

 CALL xios_set_domain_attr("domain", ibegin=ibegin, ni=ni, jbegin=jbegin, nj=nj)
 CALL xios_set_domain_attr("domain ", lonvalue_2d=lon, latvalue_2d=lat)

 dtime%second=3600
 CALL xios_set_timestep(dtime)

 CALL xios_close_context_definition()

 DO timestep=1,96
  CALL xios_update_calendar(timestep)
  CALL xios_send_field("field", field)
 ENDDO

 CALL xios_context_finalize()
 CALL xios_finalize()
END SUBROUTINE hello_world
```

Initialise XIOS and one context

Define domain

Set time step to 1 hour

End of context definition
No more modification to the context

Enter the time loop

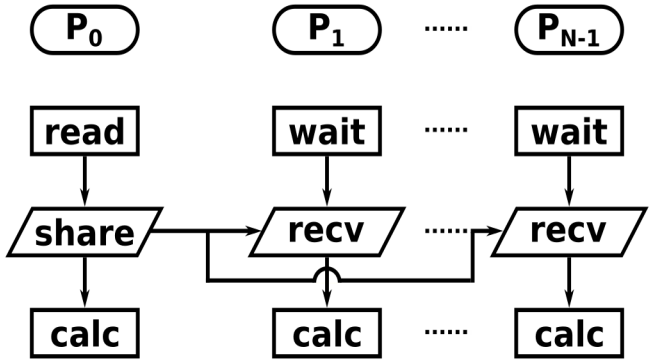
Free the context and quit XIOS

Hands-on 1

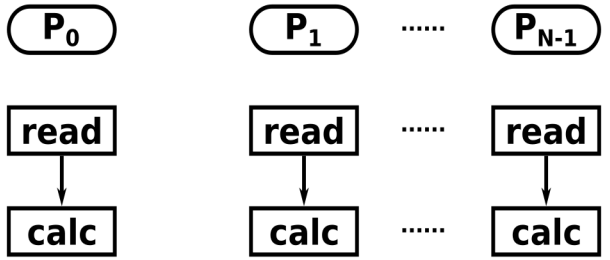
from XIOS TUTORIAL : CEA/LSCE – IPSL

TM5-MP-XIOS interface

- Default
  - Replace the TM5-MP code that reads meteo in first process with calls to XIOS
  - Difficult to get it running
  - Third attempt worked: single file – single context



- Parallel:
  - Replace the TM5-MP code that reads meteo in parallel with calls to XIOS
  - Makes use of the deprecated option with_parallel_io_meteo
  - Switches between default and parallel reading in tmm_mf_tm5_nc.F90
  - “Easy” to get it running after Default interface



TM5-MP-XIOS interface – major issues

- Documentation, documentation, documentation
- Time can only go forward, which is problematic for the adjoint
  - For a model time interval, the data from a file is first read for the start of the interval, then for the end of the interval
  - Repeated for the same interval for subsequent meteo fields
  - In view of reading data, time does something like start – end – start – end – start …, which XIOS calendar can't deal with
  - For adjoint meteo data: reverse time dimension with "ncpdq -a –time …"
  - XIOS doesn't actually read the time, it just calculates it but doesn't check it against value in file
- Meteorological data must be in a single file for XIOS to read it
  - Concatenate meteo files using ncrcat
  - Add timevalue dataset in a format that XIOS understands (i.e. don't use only netCDF dimensions, always add a dataset with the same name!)
  - Lots of data: the normal daily files, the concatenated files and the adjoint files (last two: 72GB/month)

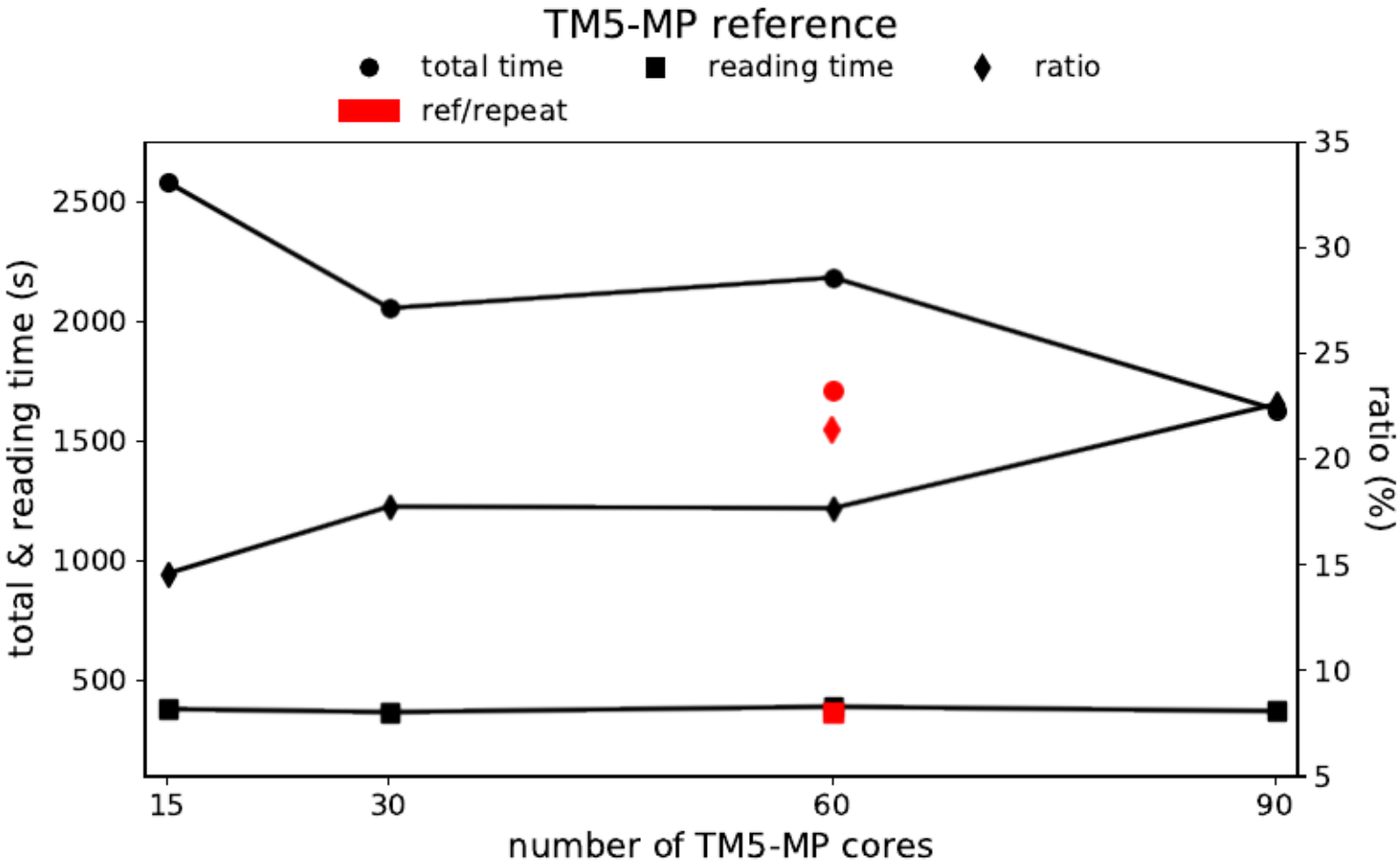
Default interface – minor issues

- Error messages are not specific enough, and XIOS function lack a status variable
- Specifying a timestep is obligatory, also for constant fields.
- The function xios_recv_field(...) accepts double precision only, e.g. no integers
- No dimension check on arrays passed to XIOS. So [lon X lat] or [lat X lon] are both possible…
- Can’t read an axis without corresponding dataset, although this is possible in netCDF
- Some subroutines are defined, but always crash if you try to call them (i.e. xios_get_timestep(...))
- It’s unclear what communicator XIOS uses, makes debugging very difficult

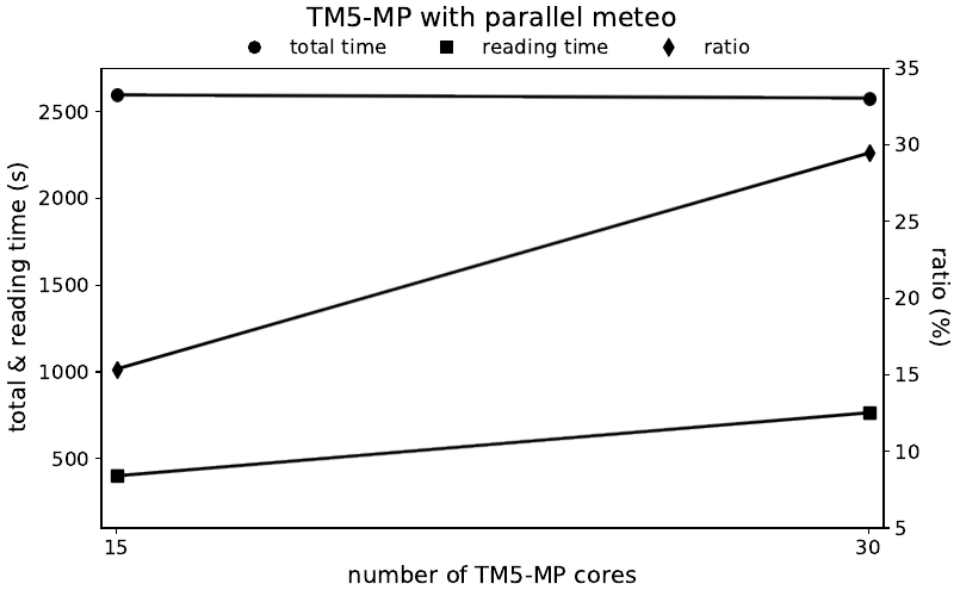
Test runs

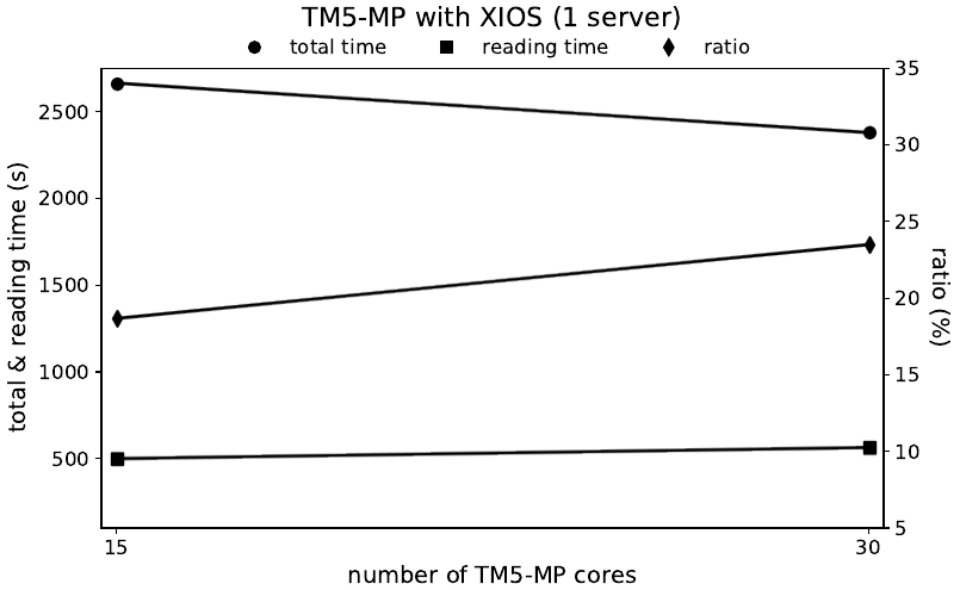
- Dutch national supercomputer Snellius
  - https://servicedesk.surf.nl/wiki/display/WIKI/Snellius
  - 1 node = 128 cores with 224GiB memory, allocation in ¼ node steps
  - Exclusive use can be requested explicitly, or by requesting all node memory
- 1 month
  - July 2015
  - ERA5 meteo on glb100x100, ml137
  - Model on glb100x100, tropo34
  - 15, 30, 60, and 90 cores for TM5-MP
  - Varying number of XIOS servers
- XIOS configuration settings
  - Mostly default
  - Minimum buffer size set to 1MB (as opposed to automatic)

Reference

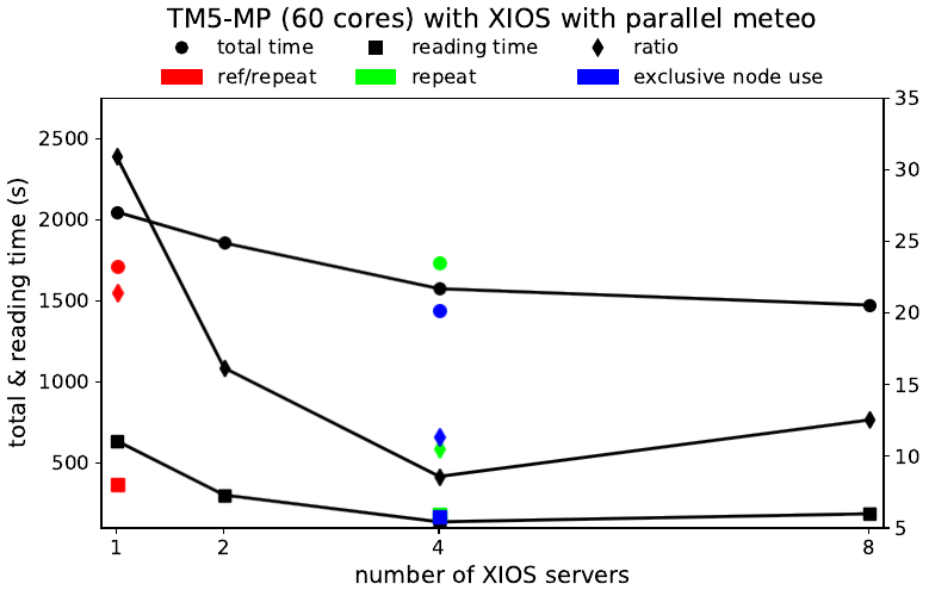

TM5-MP reference
total time
reading time
ratio
ref/repeat
total & reading time (s)
ratio (%)
number of TM5-MP cores
2500
2000
1500
1000
500
35
30
25
20
15
10
5
15
30
60
90

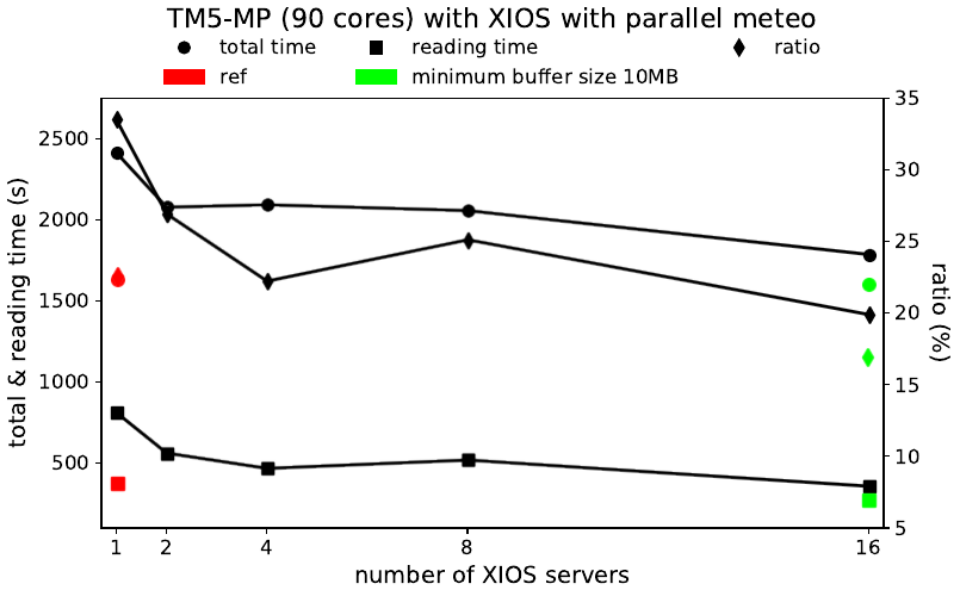
TM5-MP up to 30 cores only...





TM5-MP on 60 and 90 cores





Profiling results for TM5-MP running on 60 cores with 4 XIOS servers

	time (s)	% parent	% root	rank
root	1705,18			
step start	12,95	0,76	0,76	
read meteo	12,11	93,51	0,71	
rest	0,84	6,49	0,05	
step init	1304,81	76,52	76,52	
step init check cfl	28,78	2,21	1,69	8
step init set mass	753,57	57,75	44,19	
read meteo	149,87	19,89	8,79	
other	603,69	80,11	35,40	1
step init others	282,25	21,63	16,55	
read meteo	259,01	91,77	15,19	
other	23,23	8,23	1,36	
step init proc update	238,94	18,31	14,01	
update_kzz_read	237,20	99,27	13,91	3
other	1,74	0,73	0,10	
rest	1,27	0,10	0,07	
step run	375,52	22,02	22,02	
advectx	176,15	46,91	10,33	4
advecty	60,71	16,17	3,56	6
advectz	35,20	9,37	2,06	7
vertical	63,80	16,99	3,74	5
chemistry	8,28	2,20	0,49	
user_output	27,28	7,26	1,60	9
rest	3,57	0,95	0,21	
read meteo total	364,49	21,38	21,38	2

	time A (s)	time B (s)	time C (s)	Mean time (s)	% parent	% root	rank
root	1569,48	1726,37	1434,59	1576,81			
step start	12,09	13,59	11,9	12,53	0,79	0,79	
read meteo	11,52	12,9	11,3	11,91	95,05	0,76	
rest	0,58	0,69	0,6	0,62	4,98	0,04	
step init	1094,34	1234,81	1060,62	1129,92	71,66	71,66	
step init check cfl	67,58	67,7	30,31	55,20	4,88	3,50	7
step init set mass	669,21	724,98	664,07	686,09	60,72	43,51	
read meteo	50,69	60,24	53,37	54,77	7,98	3,47	
other	618,52	664,74	610,71	631,32	92,02	40,04	1
step init others	96,92	133,85	116,93	115,90	10,26	7,35	
read meteo	80,98	117,52	109,34	102,61	88,54	6,51	
other	15,94	16,33	7,59	13,29	11,46	0,84	
step init proc update	260,3	307,92	248,79	272,34	24,10	17,27	
update_kzz_read	258,75	306,27	247,41	270,81	99,44	17,17	2
other	1,55	1,66	1,38	1,53	0,56	0,10	
rest	0,34	0,35	0,53	0,41	0,04	0,03	
step run	456,52	465,49	352,68	424,90	26,95	26,95	
advectx	228,31	225,79	167,2	207,10	48,74	13,13	3
advecty	72,18	70,88	63,31	68,79	16,19	4,36	5
advectz	52,95	53,66	35,3	47,30	11,13	3,00	8
vertical	62,18	62,14	54,03	59,45	13,99	3,77	6
chemistry	9,74	10,15	7,85	9,25	2,18	0,59	
user_output	28,93	40,55	23,33	30,94	7,28	1,96	9
rest	2,24	2,12	1,58	1,98	0,47	0,13	
read meteo total	134,33	180,92	162,09	159,11	10,09	10,09	4

Conclusion and outlook

- Reading meteorological data with 4 XIOS servers in parallel mode is more than twice as fast as the default.
- At the moment it will take too much effort to fully implement XIOS in TM5-MP in view of the expected gain in total wall-time
  - Other processes also take a lot of time (e.g. the “other” timer on the previous slide)
  - Demand of all data to be in a single file
  - Reversal of time dimension

- Speed-up:
  - Physical parallelization (Pandey et al., 2022). Divide model period into blocks that run concurrently. Might become too resource intensive: total number of cores = number of blocks X (60+4)
  - GPU programming. Requires careful analysis of TM5-MP code. Does anyone have experience with this?
  - If significant speed-up is achieved, further integration of TM5-MP and XIOS may be considered